Modeling uncertainty and inaccuracy on data from crowdsourcing platforms: MONITOR

Constance Thierry, Jean-Christophe Dubois, Yolande Le Gall, Arnaud Martin

Modeling uncertainty and inaccuracy on data from crowdsourcing platforms: MONITOR

Constance Thierry, Jean-Christophe Dubois, Yolande Le Gall, Arnaud Martin
Univ Rennes, CNRS, IRISA
Université de Rennes 1, Lannion France
firstName.lastName@irisa.fr

***Abstract*—Crowdsourcing is characterized by the externalization of tasks to a crowd of workers. In some platforms the tasks are easy, open access and remunerated by micropayment. The crowd is very diversified due to the simplicity of the tasks, but the payment can attract malicious workers. It is essential to identify these malicious workers in order not to consider their answers. In addition, not all workers have the same qualification for a task, so it might be interesting to give more weight to those with more qualifications. In this paper we propose a new method for characterizing the profile of contributors and aggregating answers using the theory of belief functions to estimate uncertain and imprecise answers. In order to evaluate the contributor profile we consider both his qualification for the task and his behaviour during its achievement thanks to his reflection.**

***Index Terms*—Reasoning under uncertainty, Theory of belief functions, Crowdsourcing**

I. Introduction

The term crowdsourcing was introduced in [1] and defined as the outsourcing of a task to a crowd of contributors. The realized tasks on the crowdsourcing platforms are very diversified, going from micro-tasks to complex-tasks. Burger-Helmchen and Pénin [2] identified three platform types: inventive activity (complex task achievement), content (content input), routine activity (micro-task achievement). This paper focuses on the routine activity platforms where contributors perform microtasks for a micropayment. Issues related to contributors profile, answers aggregation and questions ergonomic appear in such a context.

Into routine activity platforms, the crowd is significant and diversified as shown by the demographics studies [3]–[6], including a diversity into the contributor profiles. Indeed, some contributors are more qualified to achieve the task. Moreover, not all the workers have the same conscientiousness performing the task. Generally, contributors pay attention to the task, but there is no doubt that some of them are more attracted to the prospect of easy earn compensation and therefore respond quickly and randomly. We call them spammers. Due to this diversity of contributors' profiles, the crowdsourced data is noisy and unreliable, which is a first problem on the crowdsourcing platforms as it impacts the quality of aggregate data. In particular, human contributions of uneven quality, from a heterogeneous crowd, reveal imperfections that are difficult to take into account in the decision-making process. They are inherent in any subjective assessment, may be amplified by a lack of expertise or seriousness.

Furthermore, in routine activity platforms, tasks often consist of multiple choice questionnaires. The contributor is expected to choose only one of the proposed answers. The ergonomics of this task is problematic. Indeed, if the contributor hesitates between several possible answers, he has nevertheless to choose one of them, which can eventually lead to a random response, a loss of confidence in his ability and to introduce noise into the data. Providing the possibility to be imprecise in case of doubt is an evolution in data collection that should improve the quality of data and strengthen the contributor's confidence in his answer. Taking into account data imperfections in order to model and integrate them into the information fusion process aims to facilitate optimal decision-making.

The method commonly used in the crowdsourcing platforms for aggregating the answers is the majority voting (MV) which consists of choosing the answer specified by the largest number of contributors. This method is easy to implement but has the disadvantage to do not consider the uncertainty of the contributor's answers and is thereby not robust to the spammers. Honeypot [7] uses Gold standards to identify the spammers and performs a pre-selection before the MV allows more credit to be given to aggregate data. However it is not always possible to have Gold standards for some crowdsourcing campaigns. An alternative method is the Expectation-Maximisation (EM) algorithm used by Dawid and Skene [8]. EM is also used in [9] and [10] to estimate the sensitivity and specificity of a contributor and aggregate the crowdsourcing campaign answers. Various studies [9]–[13] agree that EM outperforms the MV. Nevertheless, EM does not consider the contributor behaviour as a key element of his profile. Indeed, a lucky spammer can give the right answer to a question but he does it faster than a serious contributor. In addition, these aggregation methods are not in agreement with a possible imprecision of contributions. According to Smets [14], the fuzzy sets theory models imprecision, which is notably applied to crowdsourcing in [15]. However, fuzzy sets do not model uncertainty.

As mentioned above, crowdsourcing approaches raise several questions. In this paper, we consider both the problem of the contributor's profile and the imprecise answers. To do this we propose an innovative method, which extends our previous work presented in [16], to MOdel uNcertainty and Inaccuracy on daTa from crOwdsourcing platfoRms (MONITOR). MONI-

TOR estimates the contributor's profile by taking into account not only his qualification but also his behaviour. The main contributions of this work are:

- The estimation of the contributor's profile thanks to his knowledge and behaviour.
- The answers modeling and aggregation by the theory of belief functions.

The rest of the paper is organized as follows. In Section II the theory of belief functions is presented, then the related works on the use of this theory in crowdsourcing are reviewed Section III. The proposed model for the profile estimation MONITOR is introduced Section IV. The data sets used for the tests and the validation are described in Section V. Finally, Section VI concludes this paper.

II. Background Theory

The theory of belief functions also known as Dempster-Shafer theory has been introduced by Dempster [17] and formalized by Shafer [18] as a theory of evidence. This theory is a generalization of the fuzzy and probabilistic approaches, it allows the modeling of uncertainty and imprecision of imperfect sources. Let $\Omega = \{\omega_0, .., \omega_n\}$ a set of classes/hypothesis ω_i which are exclusive and exhaustive, Ω is named frame of discernment. In the context of crowdsourcing, a contributor c is a source of information and the frame of discernment is composed of the proposed answers for a question q. A mass function is defined for a contributor c on a question q by the function $m^{\Omega}_{cq} : 2^{\Omega} \rightarrow [0, 1]$ such that:

$$\sum_{X \in 2^{\Omega}} m^{\Omega}_{cq}(X) = 1 \tag{1}$$

Let $X \in 2^{\Omega}$, the mass $m^{\Omega}_{cq}(X)$ characterizes the belief of the contributor c into the answer X at the question q. If $m^{\Omega}_{cq}(X) > 0$ then X is called focal element. The empty set $\emptyset$ symbolizes the world openness, in the case of normalised belief function, the mass value on the empty set is null. The set Ω symbolizes the ignorance, a mass value of 1 on this set means that the contributor totally ignores which answer could be the good one. Commonly a belief function such as for one set $X \in 2^{\Omega}, m^{\Omega}_{cq}(X) = 1$ is called categorical (or logical) mass function; the contributor is absolutely certain that the answer is X, and his answer can be imprecise. Another specific belief function is the simple mass function (X^{α}):

$$\begin{cases} m^{\Omega}_{cq}(X) = \alpha \text{ with } X \in 2^{\Omega} \setminus \Omega \\ m^{\Omega}_{cq}(\Omega) = 1 - \alpha \end{cases} \tag{2}$$

The contributor is not certain of the answer X, he believes in it but not completely, once again X can be imprecise.

In case of doubt on the reliability of a source c, a discounting degree $\alpha \in [0, 1]$ can model the reliability of c.

$$\begin{aligned} m^{\Omega,\alpha}_{cq}(X) &= \alpha_{cq} m^{\Omega}_{cq}(X), \forall X \in 2^{\Omega} \setminus \Omega \\ m^{\Omega,\alpha}_{cq}(\Omega) &= 1 - \alpha_{cq}(1 - m^{\Omega}_{cq}(\Omega)) \end{aligned} \tag{3}$$

The parameter α equal to zero implies that c is absolutely not reliable at all and the mass value is affected to Ω, concluding to a total ignorance. An advantage of discounting process is that it reduces the conflict that occurs during the combination.

Many combination operators exist in the theory of belief functions for information fusion from S sources [19]–[22]. We present in this article the conjunctive combination operator which is the most common and the Yager conjunctive combination operator given by equation (5). These operators require that the sources must be reliable, distinct and independent.

$$\begin{aligned} &\text{Let } X, Y_1, ..., Y_S \in 2^{\Omega} \\ &m^{\Omega}_{Conj}(X) = \sum_{Y_1 \cap ... \cap Y_S = X} \prod_{c=1}^{S} m^{\Omega}_{cq}(Y_c) \end{aligned} \tag{4}$$

The conjunctive combination operator (Equation (4)) decreases the imprecision on the focal sets and strengthen the belief in the concordant sets between the different information sources c. The mass $m^{\Omega}_{Conj}(\emptyset)$ is called global conflict of the combination, when the sources are in conflict, this mass in non-null. To deal with this problem and stay in a closed world, Yager [22] interpreted the global conflict as the total ignorance and proposed, for $X \in 2^{\Omega}$, the rule given by:

$$\begin{aligned} m^{\Omega}_{Y}(X) &= m^{\Omega}_{Conj}(X), X \neq \emptyset, X \neq \Omega \\ m^{\Omega}_{Y}(\Omega) &= m^{\Omega}_{Conj}(\Omega) + m^{\Omega}_{Conj}(\emptyset) \end{aligned} \tag{5}$$

The combination is always done on the same frame of discernment. If a combination of information sources on different frames of discernment is expected, a vacuous extension should be done before the combination:

$$m^{\Omega \uparrow \Omega \times \Theta}(B) = \begin{cases} m^{\Omega}(A) \text{ if } B = A \times \Theta, \forall A \subset \Omega \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

Finally to make a decision on the elements of the frame of discernment, the mass function is transformed into the pignistic probability:

$$betP(X) = \sum_{Y \in 2^{\Omega}, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m^{\Omega}(Y)}{1 - m^{\Omega}(\emptyset)} \tag{7}$$

Element $\omega_i \in \Omega$ for which the maximum probability is obtained $betP(\omega_i) = \max_{\omega \in \Omega} betP(\omega)$, is selected.

The theory of belief functions is very interesting for modeling the uncertainty and imprecision of the crowdsourcing answers, considering the contributors as imperfect sources of information. Some authors use this theory in crowdsourcing, their work is presented in the following section.

III. Related Work

We differentiate two types of approaches for using belief functions in a crowdsourcing context. The first one uses the Gold standard and considers precise answers, and the latter does not use any and allows the contributor to be imprecise in his answers.

A. With Gold standard and without imprecise answer

In the approach given in [23] Gold data is used to build a reference oriented graph, calculated using expected theoretical ratings. Another graph is built from the contributors' answers and then compared to the expected graph thanks to the theory of belief functions. The goal of the study is to measure the contributor expertise by computing the distance between the graphs to differentiate the experts "E" from the non-experts "NE" consequently the frame of discernment is $\Omega = \{E, NE\}$. The weakness of this method is that it relies entirely on the Gold data. That is why the approach of [24] is interesting as it uses Gold data but the evaluation of the contributor's profile is not based solely on these data. The authors consider three profiles of contributors: "Expert", "Good" and "Bad". To start, measures are computed using Gold data, MV and logarithmic distance. Then a k-means classification is applied to estimate the profile of contributors. Finally, mass functions are assigned to the contributor's answers according to their profile: categorical mass functions for the "Expert", simple mass functions for the "Good" and the ignorance for the "Bad" one. In contrast to these approaches, those introduced in the following paragraph do not use Gold data. They are therefore less constrained and offer to the contributor the possibility of being imprecise in his answers.

B. Without Gold standard and with imprecise answer

The Cascad method proposed by [25] can be broken down into three stages. First, a qualification test is realized before the task execution to determine the contributor's profile: ignorant, little competent, averagely competent, competent, expert. To this profile is associated an imprecision degree ranging from 1 (expert) to 0 (ignorant). Secondly, during the campaign when a contributor answers question, he has to provide his uncertainty degree which is use to compute a mass function for the answer, discounted by the contributor imprecision degree. In the final stage, the discounting mass functions are combined by a cascading method based on the Dempster rule. Cascade is compared in [25] to the MV and an EM algorithm of [8] for the answer aggregation which shows that the belief functions' based method offers better results than EM and the MV.

Ben Rjab *et al.* [26] model the possible imprecision of the contributor by belief functions and identify experts without Gold data. To do that they compute exactitude (IE_c) and precision (IP_c) degrees given by the equations:

$$\begin{cases} IE_c = 1 - \dfrac{1}{|E_{Q_c}|} \displaystyle\sum_{q \in E_{Q_c}} d_J(m_c^{\Omega_q}, m_{E_C \setminus c}^{\Omega_q}) \\ m_{E_C \setminus c}^{\Omega_q}(X) = \dfrac{1}{|E_C| - 1} \displaystyle\sum_{j \in E_C \setminus c} m_j^{\Omega_q}(X) \end{cases} \tag{8}$$

$$\begin{cases} IP_c = \dfrac{1}{|E_{Q_c}|} \displaystyle\sum_{q \in E_{Q_c}} \delta_c^{\Omega_q} \\ \delta_c^{\Omega_q} = 1 - \displaystyle\sum_{X \in 2^{\Omega_q}} m_c^{\Omega_q}(X) \dfrac{log_2(|X|)}{log_2(|\Omega_q|)} \end{cases} \tag{9}$$

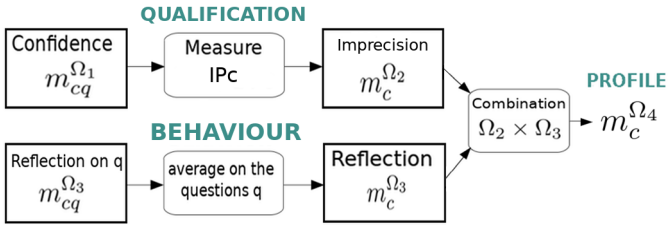

Fig. 1. MONITOR schemed

In equations (8) and (9), E_C is the contributor set, E_{Q_c} the set of questions the contributor answered and Ω_q the frame of discernment associate to the question q. Ben Rjab *et al.* assume that the majority of the contributors give the correct answer. The degree IE_c uses the Jousselme's distance [27] to measure the exactitude of the contributor's answer compared to the rest of the crowd, and IP_c the answers' dispersion weighted by the belief. Both degrees are weighted by a coefficient $\beta_c \in [0, 1]$:

$$GD_c = \beta_c IE_c + (1 - \beta_c) IP_c \tag{10}$$

This global degree allows the method to classify a contributor as expert, imprecise expert or ignorant. In their study, they use generated data to test and compare their approach to a probabilistic one and they obtain a better classification rate.

The IE_c degree is relevant when the majority of contributors is right, so the expertise associated with it is fair. Otherwise, if the majority is wrong but the contributor has a high degree of accuracy then his expertise will be high which is incorrect. The IP_c degree is more interesting because it does not require the assumption that the majority of contributors is right, which is why it is used by MONITOR.

IV. MONITOR: TO CHARACTERIZE WORKERS

We consider that the contributor can be imprecise in his contribution if he has a doubt between several answers. In this way, the contributor confidence in the answer should increase since he does not answer randomly. Thanks to a previous study the theory of belief functions is useful for modeling and aggregating answers in crowdsourcing platforms as it addresses both incertitude and imprecision. That is why we use it to model contributions as the contributor's *Confidence*. MONITOR, illustrated in Figure 1, includes the modeling of contributions (*Confidence*), as well as the qualification (*Imprecision*) and behaviour (*Reflection*) of the contributor to establish his profile. These notions are detailed bellow.

A. Self-confidence of contributors

The confidence of a contributor c in his answer to question q is modeled by the simple mass function $m_{cq}^{\Omega_1}$. The frame of discernment $\Omega_1 = \{\omega_1, \ldots, \omega_n\}$ is composed of the proposed answers to q. The mass α_{cq} associated with the contributor's answer is a numerical value representing the confidence he has given in his answer. The values associated with α_{cq} for the tests are provided in Table III.

B. Contributors' imprecision

The imprecision of a contributor represents his qualification for the task. He can be precised "P" or imprecised "NP". The belief function associated on the frame of discernment $\Omega_2 = \{P, NP\}$ is given by the equation:

$$\begin{cases} m_c^{\Omega_2}(P) = \beta * IP_c \\ m_c^{\Omega_2}(NP) = \beta * (1 - IP_c) \\ m_c^{\Omega_2}(\Omega_2) = 1 - \beta \end{cases} \quad (11)$$

In equation (11), β is a discounting factor, and the frames of discernment used in IP_c are $\Omega_q = \Omega_1$ in equation (9).

C. Contributors' reflection

To study the behaviour of a contributor, MONITOR refers to the Big Five model, also called the OCEAN model, proposed by Goldberg [29], [30] to characterize an individual's personality. Personality traits are proposed in the model are:

- Openness to the experience: A person's tendency to be open to experiences, whatever their nature. It can be characterized by the curiosity and the imagination. In crowdsourcing it can be translated by the willingness to realize the experience.
- Conscientiousness: Person who practices self-discipline, compliance and prefers organization than spontaneity. In crowdsourcing platforms objective of the conscientious contributors is to perform the task correctly, respecting the instructions.
- Extraversion: Extroverted people like to be surrounded, they are full of energy and often feel optimistic. This personality trait is essential in team work.
- Agreeableness: People who are cooperative, look for social harmony. It is persons who have qualities required for team work and be attentive in their jobs.
- Neuroticism: The emotions of people with a high degree of neuroticism can affect with their ability to reason, make decisions and cope with stressful situations.

In [28], the authors consider that personality traits have an impact on the employees performance which are: responsibility for risk, quality of work, discipline and attention, cooperation between colleagues, responsibility for results. Among these performances, quality of work, discipline and attention are key elements in a crowdsourcing work, so these personality traits have an impact on the quality of the results of a crowdsourcing campaign. This is consistent with the study of [5] who introduced this model in the context of crowdsourcing to determine the relationship between personality traits and response quality. They concluded that contributor's open minded and conscientiousness are positively correlated to the result precision.

A contributor's behaviour is estimated to be related to his personality traits so that the contributor consciousness is shaped by his reflection. The more time a contributor takes think about the task the more conscientious he is. On the contrary, a short period of reflection can have two different meanings. Either the contributor is not conscientious and answers quickly and randomly, or the contributor has better imprecision to achieve the task than the rest of the crowd and does not need the same amount of time to answer.

We focus here on the time taken by the contributor to give his answer. Consider the following frame of discernment $\Omega_3 = \{R, NR\}$, where "R" means that the contributor is estimated reflected and "NR" instinctive. Let an element $X \in 2^{\Omega_3}$ indicating the Reflection of the contributor c for a question q, we define the mass associated with X by:

$$m_{cq}^{\Omega_3}(X) = g(T_{cq}, T_{0q}, X) \quad (12)$$

where T_{cq} is the contributor's response time to question q and T_{0q} is a theoretical expected response time to q.

Algorithm 1 Function g(real: T_{cq},T_{0q}, character: X)

```
character reflection ← NR
real mass ← C1
real α3 ← alpha(Tcq, T0q)
if Tcq > (T0q + C2) then
    reflection ← R
end if
if X = reflection then
    mass ← α3
else if X = Ω3 then
    mass ← 1 − mass − α3
end if
return mass
```

Algorithm 2 Function $alpha$(real: T_{cq},T_{0q})

```
real α ← (Tcq − T0q)/T0q
if α ≤ 0 then
    α ← αmin
end if
if α ≥ 1 then
    α ← αmax
end if
return α
```

The function g given by algorithm 1, assigns a mass α_3 to the contributor according to his reflection; α_3 is computed thanks to algorithm 2. In the function $alpha$(real: T_{cq},T_{0q}), the computed mass function is bounded by a minimum (α_{min}) and a maximum (α_{max}) of alpha values. These limits are such that: $0 < \alpha_{min} < \alpha_{max} < 1$, in order to be sure to not obtain a categorical mass function for the reflection.

Once the mass function $m_{cq}^{\Omega_3}$ is computed for each question, the average on q of all these mass functions is calculated to obtain $m_c^{\Omega_3}$ which models the contributor's thinking throughout the crowdsourcing campaign.

D. Contributors' profile

The contributor profile is estimated on the Cartesian product of the frames of discernment $\Omega_4 = \Omega_2 \times \Omega_3$. To obtain Ω_4 the mass functions $m_c^{\Omega_2 \uparrow \Omega_4}$ and $m_c^{\Omega_3 \uparrow \Omega_4}$ are first computed, then the Yager conjunctive operator given in equation (5) is

TABLE I
CONTRIBUTOR PROFILES

Ω_3 \ Ω_2	Imprecise	Precise
Not Thoughtful	Expert	Spammer
Thoughtful	Fuzzy	Categorical

applied. To decide on the contributor profile, the mass function $m_c^{\Omega_4}$ is transformed into a pignistic probability. The profile with the highest probability is assigned to the contributor. Table I describes the profile of contributor under the frame of discernment Ω_4:

- **The spammer** only cares about remuneration, he does not pay attention to the task. He is not thoughtful in his work because he answers promptly to complete the task as quickly as possible. He is precise as it takes less time than being imprecise in the answers.
- **The categorical contributor** conscientiously performs the task, taking the necessary time, thus he is thoughtful. He is categorical in his answers and does not take the opportunity to be imprecise that is offered to him.
- **The fuzzy contributor** is imprecise and performs the task thoughtfully.
- **The expert** is more qualified for the task than the average contributor. His answers are therefore not thoughtful as they are instinctive. This is why he responds more quickly, while allowing himself to be imprecise when he feels the need.

V. EXPERIMENTAL RESULTS

This section first presents the experimental data set used and the proposed confidence modeling, then the results of the validation of the imprecision, reflection and profile of the contributors.

A. Real data set

To validate our method, tests are carried out on real data from a crowdsourcing campaign, performed on the Foule Factory platform [31], which consists of evaluating sound recordings. The quality scale proposed to contributors is as follow: bad (1), poor (2), correct (3), good (4), great (5). At the same time, the contributor is asked to give his confidence in his answer thanks to the confidence scale of Table III.

The campaign is composed of 4 HITs, each HIT including 12 audio records to rate. Among these records, the real sound quality of 5 of them is known, they are our Gold standards. The other 7 records are test data. Gold data are not used in the modelling proposed in MONITOR, they only served as a reference for model validation. The crowd that performed the task was consisted of 93 contributors, *i.e.* 93 answers for each signal to rate. In total the data set contains 4464 contributions.

Of these 4464 contributions, 965 are imprecise, representing 21.6% of the answers, a significant rate. In addition these imprecise answers were given by 70 of the 93 contributors who carried out the campaign, *i.e.* 75.3% of them. This high use of imprecision reflects a real need of the user.

TABLE II
SUMMARY OF THE CONFIDENCE OF THE CONTRIBUTOR FOR DIFFERENT DATA SETS

		nbData	Min	Avg	Max
IA	Global	965	**1.00**	**4.14**	5.00
	Aggregated by contributor	70	**3.00**	**4.09**	5.00
PA	Global	3 499	1.00	4.33	5.00
	Aggregated by contributor	93	3.39	4.32	5.00
PAI	Global	2 395	**2.00**	**4.34**	5.00
	Aggregated by contributor	70	**3.83**	**4.32**	5.00
PAP	Global	1104	1.00	4.29	5.00
	Aggregated by contributor	23	3.50	4.29	5.00
CA	Global	4464	1.00	4.29	5.00
	Aggregated by contributor	93	3.4	4.29	5.00

Table II summaries the confidence placed by the contributors in their answers according to five data sets: imprecise answers (IA), precise answers (PA), precise answers of contributors only that were imprecise (PAI), precise answered of contributors that where always precise (PAP) and the whole crowd (CA). For the fives data sets, minimum (Min), average (Avg) and maximum (Max) confidence values are provide, a global on all the answers, and another, for the answers aggregated (by mean) by contributor. The number of data in the dataset is given by the column nbData. For example, for the IA, the global data set consists of 965 confidences on answers and the aggregate data set per contributor consists of the average confidence of 70 contributors.

For all the data sets, the maximum of confidence is 5.00 (“very confident") which is the most important confidence proposed to contributors. The minimum confidence level is 1.00 except for global answers from the PAI. This is interesting, because the minimum of the aggregate average confidence per contributor is also the highest for the PAI. Contributors who have sometimes been imprecise in their answers are less confident than the others. But when these imprecise contributors give a precise answer then, they are more confident than the average. We observe it in Table II for the minimum and the average of confidence.

B. Confidence experimental modeling

The frame of discernment on the answers associated with the data is $\Omega_1 = \{bad, poor, correct, good, great\}$. The numerical values associated with contributors' confidence are presented in Table III. These values are scaled from 0 to 1 with a step of $\frac{1}{|\text{Confidence scale}|}$, i.e. here 0.25. The numerical values assigned to “Very confident" and “Not confident" have been respectively reduced and increased by 0.01 in order not to have categorical mass functions. It is desirable to maintain uncertainty on the contributor answer even if he is fully confident in his answer.

C. Aggregation

In this section comparisons with an aggregation by MV are done. Usually on crowdsourcing platforms, the contributor is asked to be precise in his answer. But in this study,

TABLE III
CONFIDENCE AND NUMERICAL VALUES ASSOCIATED

Confidence scale	α_{cq}
Very confident	0.99
Quite confident	0.75
Moderately confident	0.5
Few confident	0.25
Not confident	0.01

TABLE IV
ANSWER ERROR RATE FOR DIFFERENT VALUES OF λ

HIT \ λ	1	0.5	0.4	0.3	0.2	0.1	0	MV
0	0.6	0.4	0.4	0.4	0.2	0.2	0.2	0.6
1	0.6	0.6	0.6	0.6	0.6	0.8	0.6	0.4
2	0.4	0.6	0.6	0.6	0.4	0.2	0.2	0.4
3	0.4	0.2	0.2	0.2	0.2	0.2	0.2	0.6
All	0.5	0.45	0.45	0.45	0.35	0.35	0.3	0.5

TABLE V
QUALITY MARKS ESTIMATED BY AGGREGATION OF THE FOUR HITS

Gold data	MV	$m_q^{\Omega_1}$	$m_1^{\Omega_1}$	$m_{0.5}^{\Omega_1}$	$m_{0.2}^{\Omega_1}$	$m_0^{\Omega_1}$
1	1	1	1	1	**1**	**1,2**
2	1	1	1	1	**2**	**2**
3	4	4	4	4	**3**	**3**
4	4	4	4	4	**4**	**4**
5	4	4	4	4	**4**	**4**

the contributor is allowed to be imprecise (by selecting a maximum of two choices), and that is why his two answers are taken into account for the MV.

It is complicated to use the conjunctive operator for the answer aggregation, as there is a lot of data and the masses resulting from conjunctive aggregation are very low. In this case, it is not possible to make a decision. To address this problem, an aggregation by mean is performed to obtain $m_q^{\Omega_1}$. The results are equal to those of the MV as shown in Table V, but according to [14], increasing imprecision should decrease uncertainty. To interpret this idea, the aggregation by the average value, of belief functions for precise ($m_P^{\Omega_1}$) and imprecise ($m_{IP}^{\Omega_1}$) answers is differentiated. Then, both mass values are weighted by a coefficient $\lambda \in [0,1]$ and added:

$$m_\lambda^{\Omega_1} = \lambda * m_P^{\Omega_1} + (1-\lambda) * m_{IP}^{\Omega_1} \quad (13)$$

For $\lambda = 1$, $m_\lambda^{\Omega_1} = m_P^{\Omega_1}$ only precise answers are used, respectively $\lambda = 0$, $m_\lambda^{\Omega_1} = m_{IP}^{\Omega_1}$.

Once $m_\lambda^{\Omega_1}$ is obtained, a pignistic probability is applied to determine the answer. Different lambda coefficients are tested for the 5 Gold data of the four HITs. For each λ tested, the answer obtained is compared with the expected data. Answer error rates are given in Table IV, for HITs 0 to 3 the five answers are considered, which represents a total of 20 answers. Table IV shows that by decreasing λ, the correct answer rate increases for all HITs and more importance is then given to inaccurate answers. Moreover, the worst-case answer error rate for $\lambda = 1$, is equal to that of the MV and best-case answer rate, for $\lambda = 0$ is 20% lower than the MV.

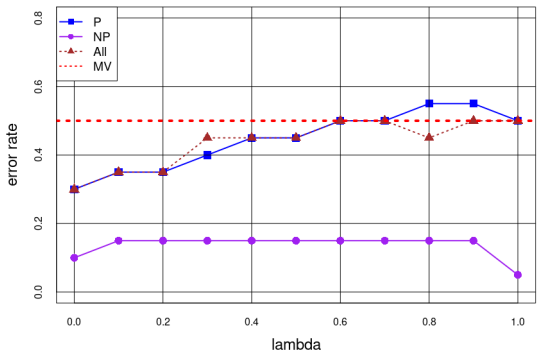

Fig. 2. Evolution of the error rate as a function of λ according contributors groups based on **precision**

The quality marks estimated by the contributors, for $\lambda \in \{1, 0.5, 0.4, 0.3\}$ are the same. They are given for $\lambda = 1$ and $\lambda = 0.5$ in Table V. In a same way, the quality marks of $\lambda = 0.2$ and $\lambda = 0.1$ are equal and given for $\lambda = 0.2$ in the same Table. The aggregation of precise answers gives the same results as MV and the same is true for $\lambda \in \{0.5, 0.4, 0.3\}$ which is consistent with their low correct answer rate. On the other hand, imprecise answers provide better estimations of marks, but do not allow to decide for the first question which mark to choose between 1 and 2. Finally, $\lambda = 0.2$ and $\lambda = 0.1$ provide a good aggregation of the answer compared to the others and deals with the indecision of imprecise answers.

Note that for all aggregations in Table V, the maximum quality given to the best quality record (5) is “good" (4). This could be explained by the difference in quality between the “good" record and the “great" one, which is so weak that it is hardly noticeable and misleading to contributors. But it could also be due to a cultural effect, contributors do not dare to give the highest quality mark when answering. This overall poor estimate on the 4 HITs accounts for 0.2.

Giving more weight to inaccurate answers significantly reduces the error rate, the introduction of this concept in crowdsourcing platforms seems to be a real asset.

D. Imprecision Evaluation

To estimate the imprecision of the contributor $\beta = 0.8$ is considered in equation (11) . For the evaluation of the imprecision the mass $m_c^{\Omega_2}$ is transformed in a pignistic probability. By this way, two groups of contributors are done, one of precise contributors “P" and the other one of imprecise contributors “NP". The imprecise contributors’ group is only 1.08% of the crowd. This can be explained by the fact that it is not common for contributors to be inaccurate in crowdsourcing platforms and that it is counter-intuitive to them.

In Figure 2, the answers are aggregated by group of contributors, P and NP, thanks to the equation (13). The error rate of the imprecise contributor exceeds that of others. Moreover, since is error rate is less than 0.2%, this means that the NP can be used to determine the accurate quality of the most complex record. For $\lambda < 0.6$ the error rate of the precise contributors is lower than that of the MV. Note that for $\lambda = 0.8$ and 0.9

the error rate is greater than the MV error rate, but this is not the case for crowd aggregation which means that answers of imprecise contributors have a strong impact on the global aggregation. In adition, as observed in Table IV for the all HITs, the error rate increases with lambda.

By selecting the contributors most likely to be imprecise in aggregating responses, the error rate obtained is the lowest of all. This imprecision, as we have announced, is therefore really beneficial in the context of crowdsourcing.

E. Reflection Evaluation

Since there is no reference time for completing the task, the time (in seconds) T_{0q} for each question is the time of the audio record. In algorithm 1, constant $C_1 = 0.15$ and $C_2 = 10$, this way the contributor must listen completely to the record and spend at least 10 seconds to answer the question. In algorithm 2, the constants α_{min} and α_{max} are respectively equal to 0.01 and 0.99.

To estimate the theoretical thought of a contributor, an estimation of the expected time of the campaign T_{th} is computed. As it consisted of listening sound records, the recording times T_{rec_q} are added with a constant of 10 seconds corresponding to the expected answer time of the contributor: $T_{th} = \sum_q T_{rec_q} + 10$. Then we compute the realisation time of the campaign by the contributor T_c by summing his answer times to each questions: $T_c = \sum_q T_{cq}$. If $T_c \geq T_{th}$ the contributor is estimated theoretically thoughtful. The estimation by MONITOR of the contributor thought, for the tests on the Gold data, is done by transforming the mass function $m_c^{\Omega_3}$ into a probabilistic probability.

MONITOR classifies 65.6% of contributors as thoughtful “R", which seems correct as we expected the majority of contributors to take their time to complete the campaign. For 3.2% of contributors it is not possible to decide between thoughtful or not, so we estimate that for a pignistic probability higher than or equal to 0.5 the contributor is estimated thoughtful. The remaining 31.2% are not thoughtful “NR". That seems to be a lot, but in this class there are both spammers and experts, that explains this percentage. Compared to the theoretical estimation of the thought of contributors to that of MONITOR, we found a CCR of 51.6%, which is correct once again as Gold data are not used in MONITOR.

Consider now the aggregation of the responses of thoughtful and not thoughtful contributors in Figure 3. The error rate of NR contributors converges quickly towards the MV one. Since the error rate is higher for NR than the error rate for the entire crowd, this suggests that there are more spammers in the crowd than experts. The R contributors performed globally better than the NR, and better than the crowd error rate for $\lambda < 0.8$. The differentiated aggregation of contributors according to their reflection is positive since thoughtful contributors have a lower error rate than the rest of the crowd.

The representation of the contributor’s conscientiousness through his reflection is relevant here. Indeed, as expected, thoughtful contributors are conscientious in the execution of the task, which results in a lower error rate than for others.

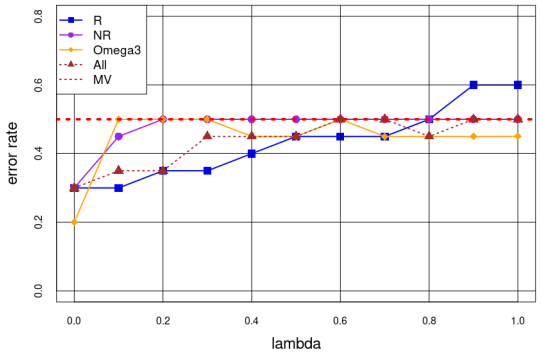

Fig. 3. Evolution of the error rate as a function of λ according contributors groups based on **reflection**

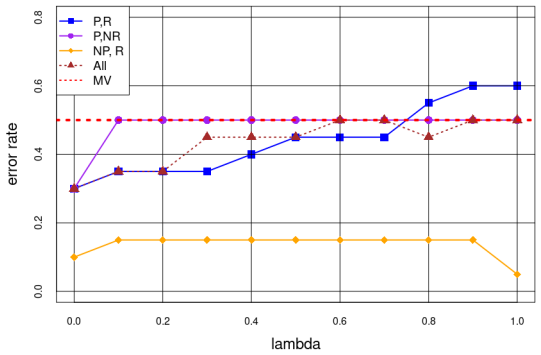

Fig. 4. Evolution of the error rate as a function of λ according contributors groups based on **profile evaluation**

F. Profile Evaluation

For profile validation, groups of contributor profiles are defined and for each group the answers are aggregated thanks to equation (13). The performed aggregations and the taken decision are made as indicated in Table V. Figure 4 compares the error rate of the answer aggregation by contributor profile to those of the MV and the aggregation of all the crowd without profile distinction. MONITOR classified 62.4% of contributors as categorical (P,R), 30.1% as spammers (P,NR), 1.1% as fuzzy (NP,R). There are no contributors classified as experts (NP,NR), who should be, in this study, a contributor having an absolute pitch. But in [32], the authors point out that having an absolute pitch is very rare and has been estimated as less than 0.01% of the general population. There should therefore be 0.93 contributors with an absolute pitch present in the crowd of this study, which explains the absence of an expert. In addition, there are some contributors (6.45%) for whom the maximum of pignistic probability is given for (P,R)∪(P,NR). The hesitation between the two profiles is due to the contributor’s reflection that does not allow to decide.

Spammers (P,NR) have the worst error rate for $\lambda < 0.8$ and have the same as the MV, its poor performance is normal and was expected. Categorical contributors (P,R) perform as well if not better than the crowd as a whole, so they are more efficient than spammers and MV. The fuzzy contributor (NP,R) is the one who achieves the best performance.

Unexpectedly, he has the best error rate for $\lambda = 1$ while for the other contributor groups the error rate for this λ is the worst.

Globally, the error rate increases with λ for all groups, so giving more weight to an imprecise answer is beneficial for data aggregation. The error rates of contributors groups according to their profiles are in line with the expected results, which provides a good estimate of these profiles.

VI. Conclusion

Main issues in crowdsourcing platforms of routine activities are task ergonomics, the estimation of the contributors profile and the answers aggregation. To challenge this issues, we offer the contributor the opportunity to be imprecise in his answers when necessary, and we ask him to precise his confidence in his answers. To deal with the imprecision and the uncertainty of contributors' answers, the theory of belief functions theory is used in the proposed method: MONITOR for the estimation of the contributor profile and the answer aggregation. To test MONITOR, a crowdsourcing campaign was carried out, 75% of the contributors who achieved it have been imprecise. We observed that when contributors are imprecise their confidence is slightly lower than average, but when they are accurate, their confidence is higher than average. So offering the ability to be imprecise in crowdsourcing platforms is an advantage for the contributors confidence. Moreover, the error rate on the answer aggregation decreases as the weight given to imprecise responses increases. Therefore, the aggregation proposed in equation (13) that models imprecision thanks to the belief functions offers good results compared to the commonly used MV. Finally, considering the answers aggregation according to the MONITOR estimated profile, we observe some interesting results. The error rates obtained according to the different contributor profiles are those we expected. Unfortunately, the campaign used for the study does not allow us to observe all types of profiles defined because of the type of expertise required. In our future work we will carry out crowdsourcing campaigns that require less specific expertise.

References

[1] J. Howe, "The Rise of Crowdsourcing", Wired Magazine, 2006

[2] T. Burger-Helmchen and J. Pénin, "Crowdsourcing : définition, enjeux, typologie", Management & Avenir, vol. 41, pp. 254–269, 2011.

[3] G. Kazai, J. Kamps, N. Milic-Frayling, "An analysis of human factors and label accuracy in crowdsourcing relevance judgments", Information retrieval, vol. 16, no. 2, pp. 138–178, 2013, Springer.

[4] J. Ross, A. Zaldivar, L. Irani and B. Tomlinson, "Who are the turkers? worker demographics in amazon mechanical turk", Department of Informatics, University of California, Irvine, USA, Tech. Rep, 2009

[5] G. Kazai, J. Kamps and N. Milic-Frayling, "The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy", Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, pp. 2583–2586, 2012.

[6] G. Kazai, J. Kamps and N. Milic-Frayling, "Worker Types and Personality Traits in Crowdsourcing Relevance Labels", 20th ACM Conference on Information and Knowledge Management, CIKM, 2011.

[7] K. Lee, J. Caverlee and S. Webb, "The social honeypot project: protecting online communities from spammers",Proceedings of the 19th international conference on World wide web, pp. 1139–1140, 2010, ACM

[8] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm", Applied statistics, pp. 20–28, 1979, JSTOR.

[9] V. C. Raykar and S. Yu, "Annotation models for crowdsourced ordinal data", Journal of Machine Learning Research, 2012.

[10] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni and L. Moy, "Learning From Crowds", Journal of Machine Learning Research, 2010.

[11] J. Whitehill, T.-F. Wu, J. Bergsma, J. R. Movellan and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise", Advances in neural information processing systems, pp. 2035–2043, 2009.

[12] N. Q. V. Hung, N. T. Tam, L. N. Tran and K. Alberer, "An evaluation of aggregation techniques in crowdsourcing", International Conference on Web Information Systems Engineering, pp. 1–15, 2013, Springer

[13] F. K. Khattak and A. Salleb-Aouissi, "Quality control of crowd labeling through expert evaluation", Proceedings of the NIPS 2nd Workshop on Computational Social Science and the Wisdom of Crowds, vol. 2, 2011

[14] P. Smets, "Imperfect Information: Imprecision and Uncertainty", Uncertainty Management in Information Systems: From Needs to Solutions, pp. 225–254, A. Motra and P. Smets, Springer US, Boston, MA, 1997

[15] C. Wagner and D. T. Anderson, "Extracting meta-measures from data for fuzzy aggregation of crowd sourced information", 2012 IEEE International Conference on Fuzzy Systems, pp. 1–8, 2012, IEEE

[16] C. Thierry, J.-C. Dubois, Y. Le Gall and A. Martin, "Modélisation du profil des contributeurs dans les plateformes de crowdsourcing", 27èmes rencontres francophones sur la logique floue et ses applications, 2018.

[17] A. P. Dempster, "Upper and Lower Probabilities Induced by a Multivalued Mapping", The Annals of Mathematical Statistics, vol. 38, pp. 325–339, 1967.

[18] G. Shafer, "A mathematical theory of evidence", vol. 42, Princeton university press, 1976

[19] E. Lefèvre and Z. Elouedi, "How to preserve the conflict as an alarm in the combination of belief functions?", Decision Support Systems, vol. 56, pp. 326–333, 2013, Elsevier

[20] M. C. Florea, A.-L. Jousselme, É. Bossé and D. Grenier, "Robust combination rules for evidence theory", Information Fusion, vol. 10, no. 2, pp. 183–197, 2009, Elsevier

[21] A. Martin, "Conflict management in information fusion with belief functions", Information Quality in Information Fusion and Decision Making, pp. 79–97, 2019, Springer

[22] R. R. Yager, "On the Dempster-Shafer framework and new combination rules", Information sciences, vol. 41, no. 2, pp. 93–137, 1987, Elsevier.

[23] J.-C. Dubois, L. Gros, M. Kharoune, Y. Le Gall, A. Martin, Z. Miklós and H. Ouni, "Measuring the Expertise of Workers for Crowdsourcing Applications", Advances in Knowledge Discovery and Management, pp. 139–157, 2019, Springer

[24] L. Abassi and I. Boukhris, "A worker clustering-based approach of label aggregation under the belief function theory", Applied Intelligence, pp. 1–10, 2018, Springer.

[25] D. Koulougli, A. Hadjali, and I. Rassoul, "Handling query answering in crowdsourcing systems: A belief function-based approach", Fuzzy Information Processing Society (NAFIPS), 2016 Annual Conference of the North American, pp. 1–6, 2016, IEEE.

[26] A. B. Rjab, M. Kharoune, Z. Miklos and A. Martin, "Characterization of experts in crowdsourcing platforms", Belief Functions: Theory and Applications, vol. 9861, 2016

[27] A.-L. Jousselme, D. Grenier and É. Bossé, "A new distance between two bodies of evidence", Information fusion, vol. 2, no. 2, pp. 91–101, 2001, Elsevier.

[28] M. S. Mehmood, A. Mehmood and M. Siddique, "Personality Traits Nexus Employee's Performance: An Application of Big Five Personality Dimensions Model", Abasyn Journal of Social Sciences–Special Issue: AIC, pp. 101–119, 2016.

[29] L. R. Goldberg, "The structure of phenotypic personality traits", American psychologist, vol. 48, no. 1, 1993

[30] L. R. Goldberg, "An alternative description of personality: the big-five factor structure", Journal of personality and social psychology, vol. 59, no. 6, 1990

[31] https://www.foulefactory.com/en/

[32] A. H. Takeuchi and S. H. Hulse, "Absolute pitch.", Psychological bulletin, vol. 113, no. 2, pp. 345, American Psychological Association, 1993